Introduction to Explainable AI

Helmholtz AI @ Summer Academy II
22.09.2023

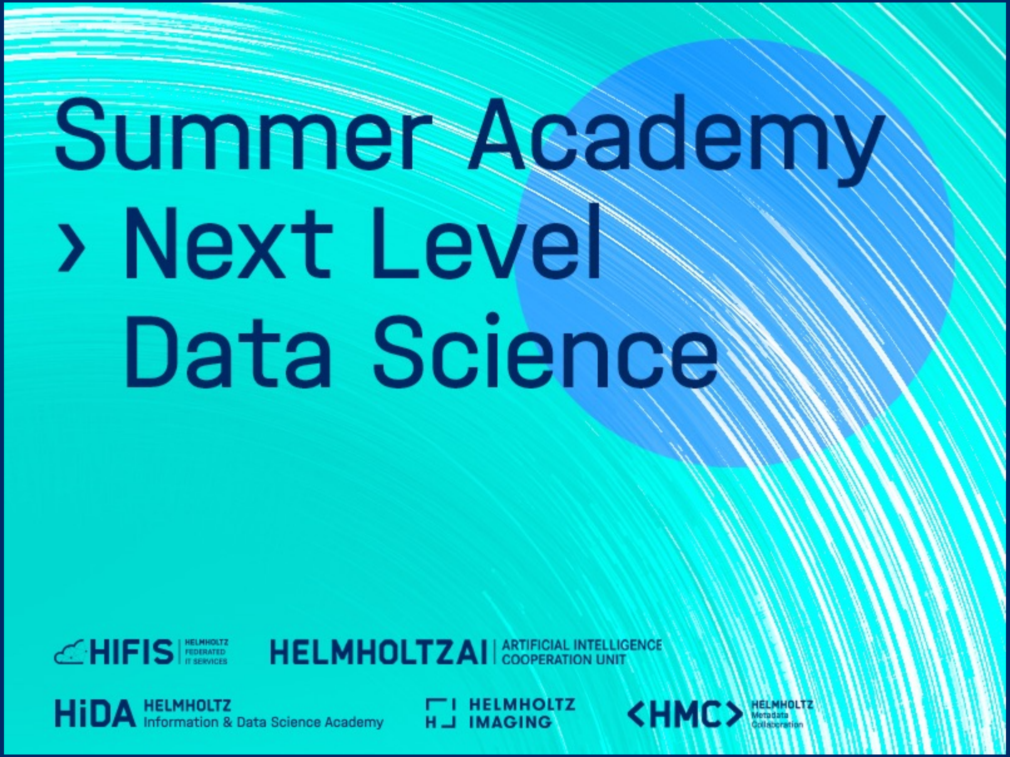
HELMHOLTZ

Incubator Summer Academy

Research for grand challenges.

- September 18- 29, 2023
- 14 course packages offered by the 5 Information & Data Science platforms
- Meet the platforms and their offers here in Gathertown!
- Exchange in our networking area!
- Please evaluate the Incubator Summer Academy! Follow this link to our feedback survey:

https://events.hifis.net/event/858/surveys/228/



HELMHOLTZ Incubator Summer Academy

Research for grand challenges.

HiDA HELMHOLTZ Information & Data Science Academy

Umbrella for 6 research schools & complementary training, networking and scouting for Centers

HELMHOLTZ AI | ARTIFICIAL INTELLIGENCE COOPERATION UNIT

Machine Learning & Artificial Intelligence

HELMHOLTZ IMAGING

Imaging techniques and image data analysis

the Helmholtz-Incubator Information & Data Science

HMC | HELMHOLTZ METADATA COLLABORATION

FAIR Research Data through enriched Metadata

HIFIS | HELMHOLTZ FEDERATED IT SERVICES

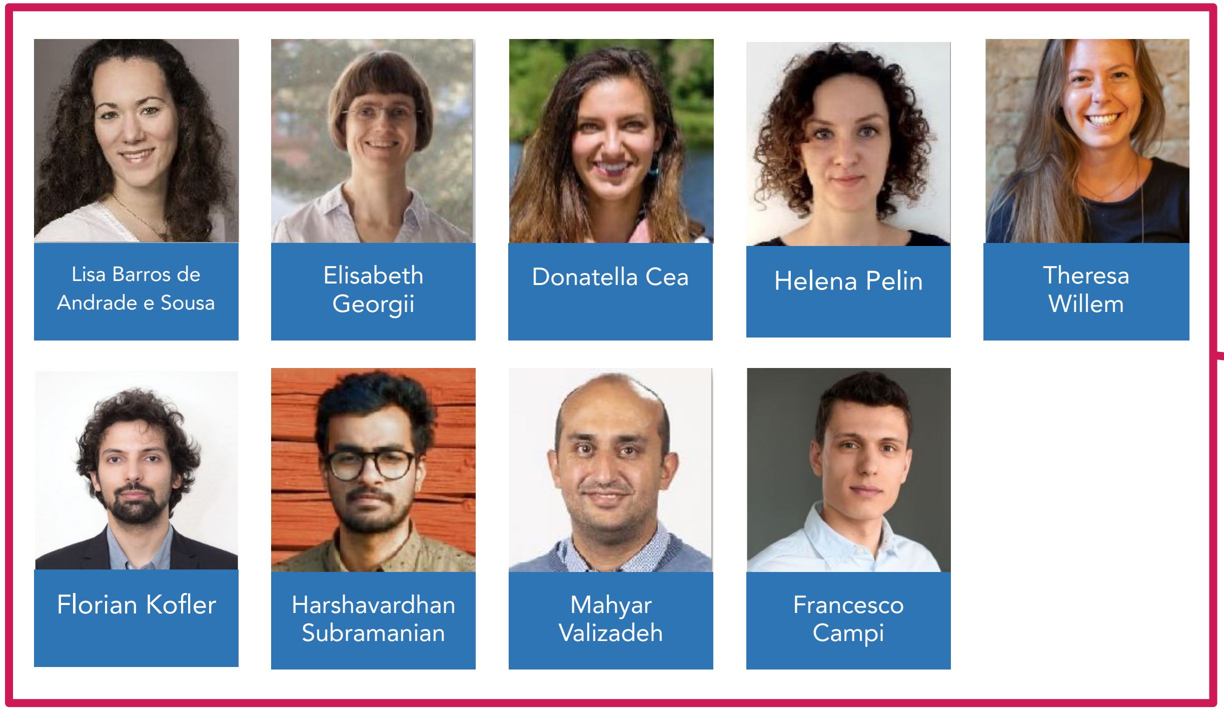
Technologies and Systems for data-based research

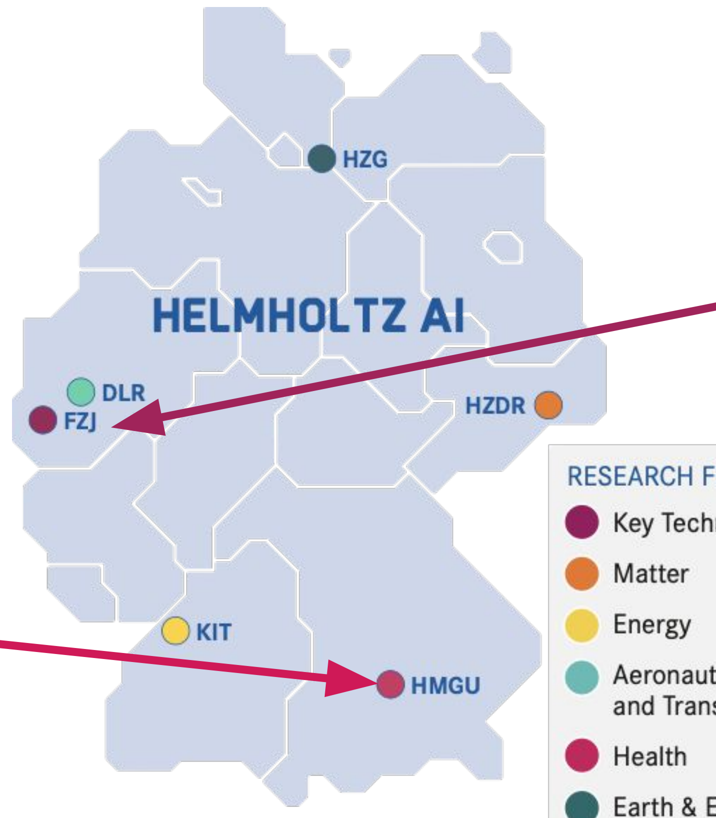
Who are we?

Helmholtz AI

WHAT IS OUR MISSION?

Maximise research impact by democratising access to AI





slido

What is your field of study?

ⓘ Start presenting to display the poll results on this slide.

Outline

Schedule and Tools

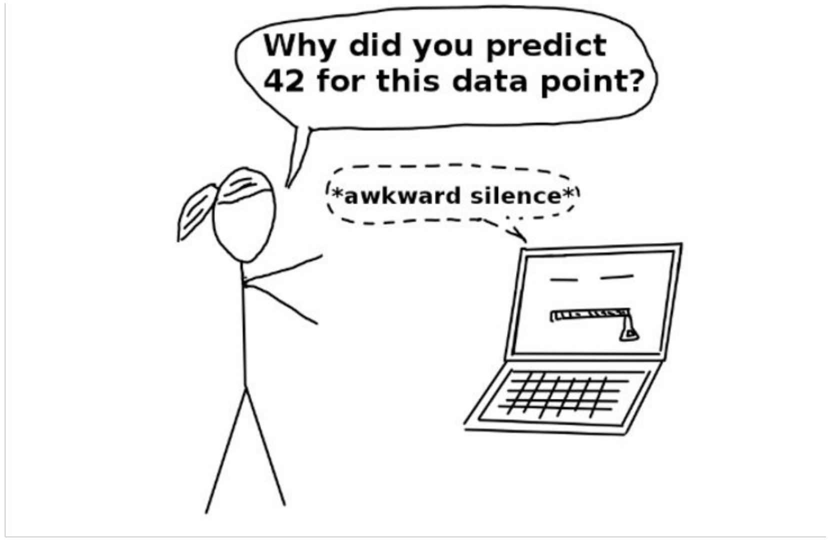
Flipped classroom approach

13.30 - 13.50	Introduction on XAI
13.50 - 15.50	XAI Model-Agnostic Methods (2 or 1 longer break when needed in individual groups)
15.50 - 16.00	Break
16.00 - 17.30	"XAI in deep learning-based image analysis" or "XAI for Random Forests"
17.30 - 17.35	Wrap-up and conclusions

Introduction

Terminology

Explainability or Interpretability?

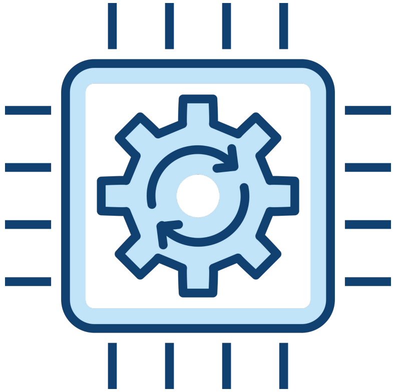


Introduction

Terminology

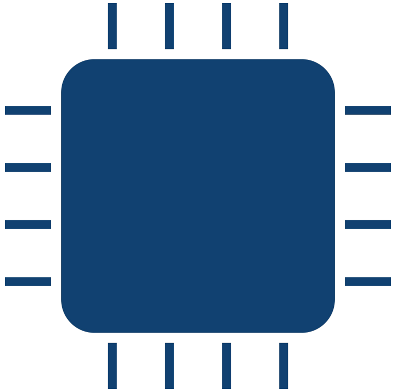
Interpretability

Understand exactly why and how the model is generating predictions by observing the inner mechanics of the AI/ML method.



Explainability

Focus on the decision-making process and try to explain the behaviour in human understandable terms.

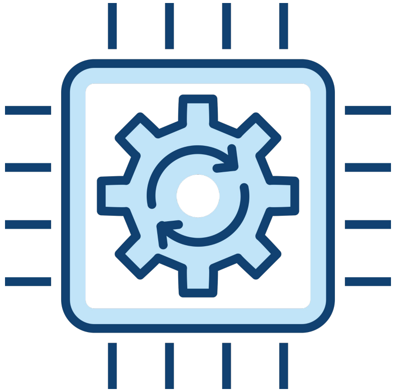


Introduction

Terminology

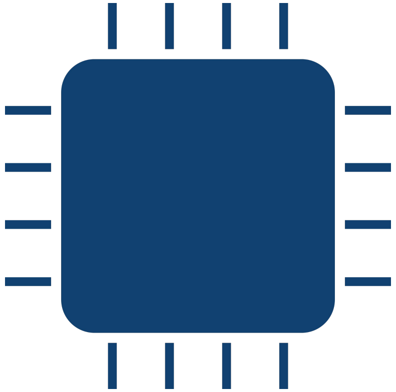
Interpretability

Understand exactly why and how the model is generating predictions by observing the inner mechanics of the AI/ML method.



Explainability

Focus on the decision-making process and try to explain the behaviour in human understandable terms.

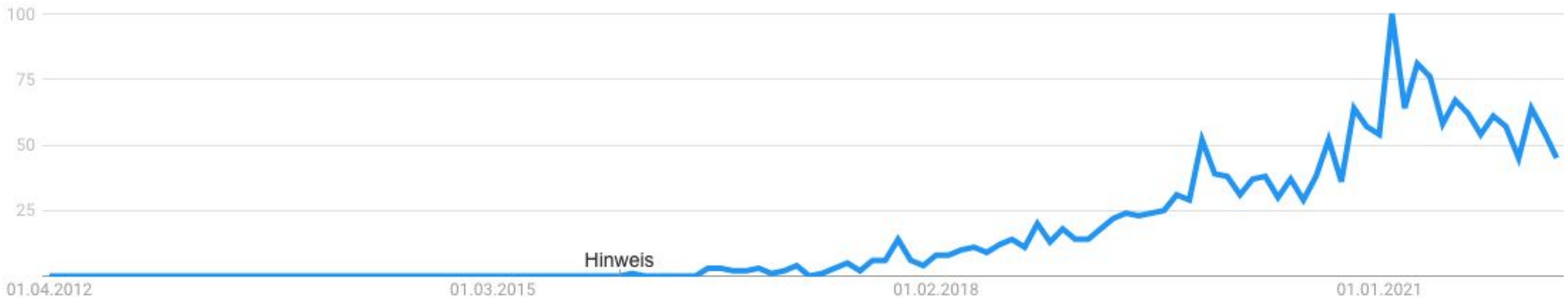


In this course, we will focus only on **eXplainable Artificial Intelligence** (XAI).

Introduction

Why is explainability important?

Google Trends Popularity Index of the term *Explainable AI* over the last ten years (2012–2022)



slido

Why is explainability important?

ⓘ Start presenting to display the poll results on this slide.

Introduction

Why is explainability important?

„The problem is that a single metric, such as classification accuracy, is an incomplete description of most real-world tasks." — (Doshi-Velez et al., 2017)

Introduction

Why is explainability important?

„The problem is that a single metric, such as classification accuracy, is an incomplete description of most real-world tasks." — (Doshi-Velez et al., 2017)

Introduction

XAI is important for technology acceptance

Introduction

XAI is important to avoid ethical issues

NEWS | 24 October 2019 | Update 26 October 2019

Millions of black people affected by racial bias in health-care algorithms

Study reveals rampant racism in decision-making software used by US hospitals — and highlights ways to correct it.

Heidi Ledford

Introduction

XAI is important for knowledge creation

What Does Deep Learning See? Insights From a Classifier Trained to Predict Contrast Enhancement Phase From CT Images

Kenneth A. Philbrick[1]
Kotaro Yoshida
Dai Inoue
Zeynettin Akkus
Timothy L. Kline
Alexander D. Weston
Panagiotis Korfiatis
Naoki Takahashi
Bradley J. Erickson

OBJECTIVE. Deep learning has shown great promise for improving medical image classification tasks. However, knowing what aspects of an image the deep learning system uses or, in a manner of speaking, sees to make its prediction is difficult.

MATERIALS AND METHODS. Within a radiologic imaging context, we investigated the utility of methods designed to identify features within images on which deep learning activates. In this study, we developed a classifier to identify contrast enhancement phase from whole-slice CT data. We then used this classifier as an easily interpretable system to explore the utility of class activation map (CAMs), gradient-weighted class activation maps (Grad-CAMs), saliency maps, guided backpropagation maps, and the saliency activation map, a novel map reported here, to identify image features the model used when performing prediction.

RESULTS. All techniques identified voxels within imaging that the classifier used. SAMs had greater specificity than did guided backpropagation maps, CAMs, and Grad-CAMs at identifying voxels within imaging that the model used to perform prediction. At shallow network layers, SAMs had greater specificity than Grad-CAMs at identifying input voxels that the layers within the model used to perform prediction.

CONCLUSION. As a whole, voxel-level visualizations and visualizations of the imaging features that activate shallow network layers are powerful techniques to identify features that deep learning models use when performing prediction.

Introduction

XAI is important to meet regulatory requirements


The Alan Turing Institute

See the full story and others like it at turing.ac.uk/partners/impact-stories

Impact story

A right to explanation

Advice from Turing researchers, urging the need for individuals to have a legally-binding right to have automated decisions made about them explained, is helping shape how the new EU general data protection regulations (GDPR) will be implemented.

Introduction

XAI is important as a defense strategy

Home › Artificial Intelligence and Soft Computing › Conference paper

Explainable AI for Inspecting Adversarial Attacks on Deep Neural Networks

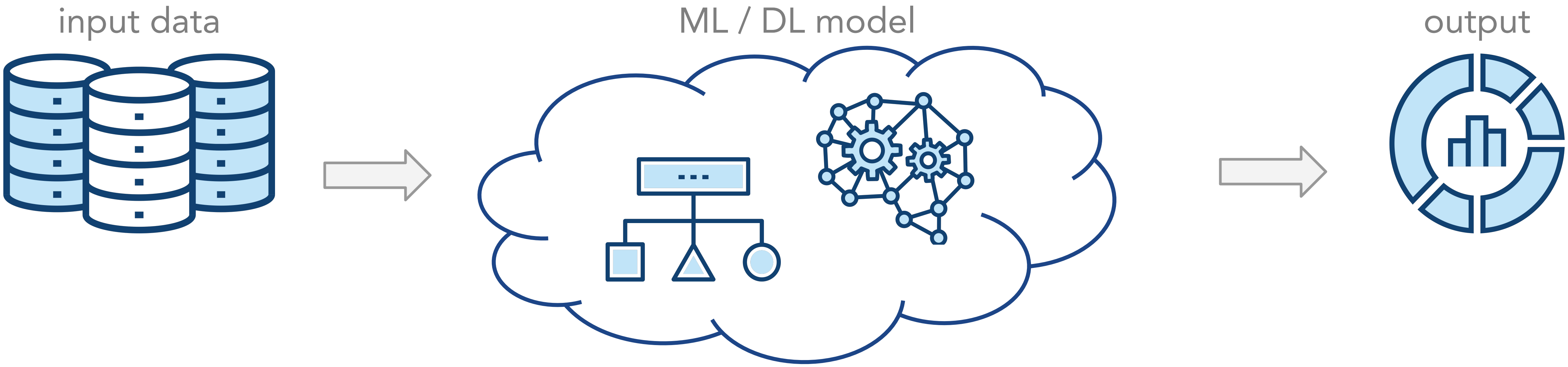
Zuzanna Klawikowska, Agnieszka Mikołajczyk & Michał Grochowski

Conference paper | First Online: 07 October 2020

Part of the Lecture Notes in Computer Science book series (LNAI,volume 12415)

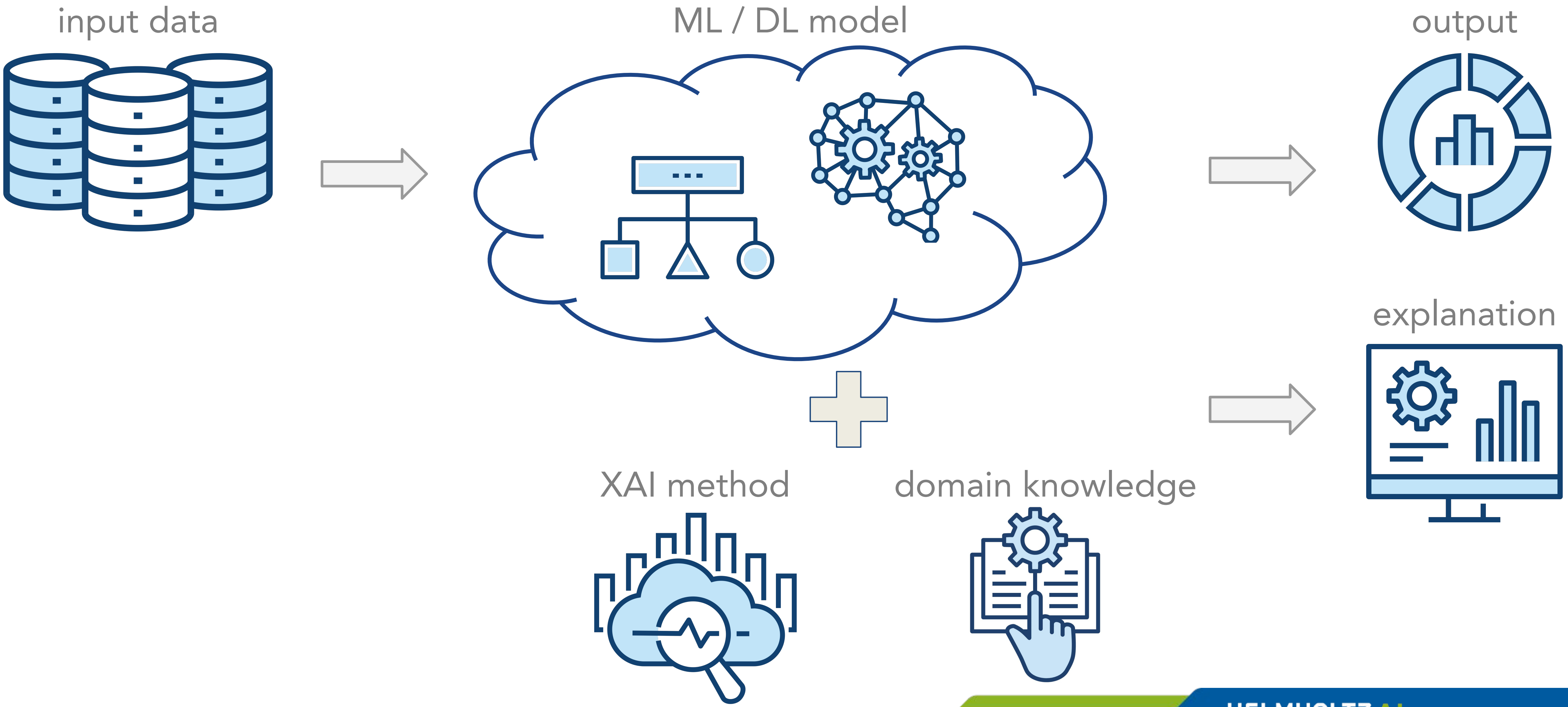
Introduction

XAI in your ML workflow



Introduction

XAI in your ML workflow

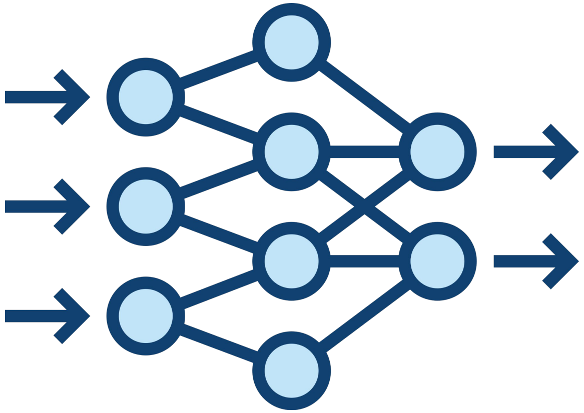


Introduction

XAI in your ML workflow

input data

ML / DL model



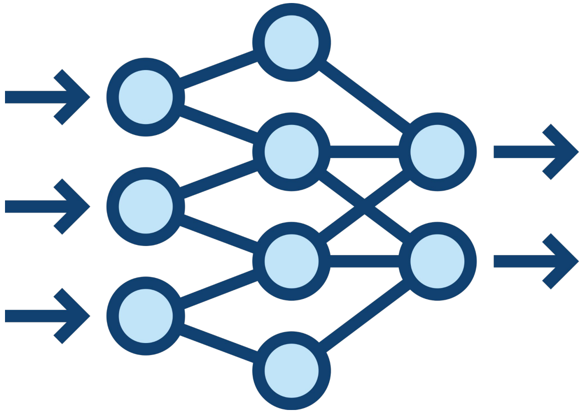
output

Introduction

XAI in your ML workflow

input data

ML / DL model



Current explanation:
This is a cat!

output

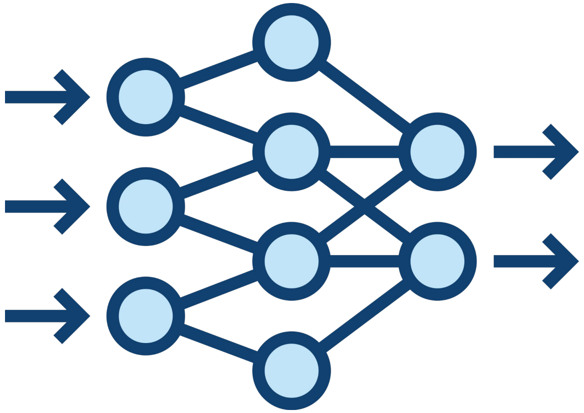
Cat

Introduction

XAI in your ML workflow

input data

ML / DL model



Current explanation:
This is a cat!

output

Cat

XAI method + domain knowledge

XAI explanation:
- it has fur, whiskers, and claws
- it has this feature

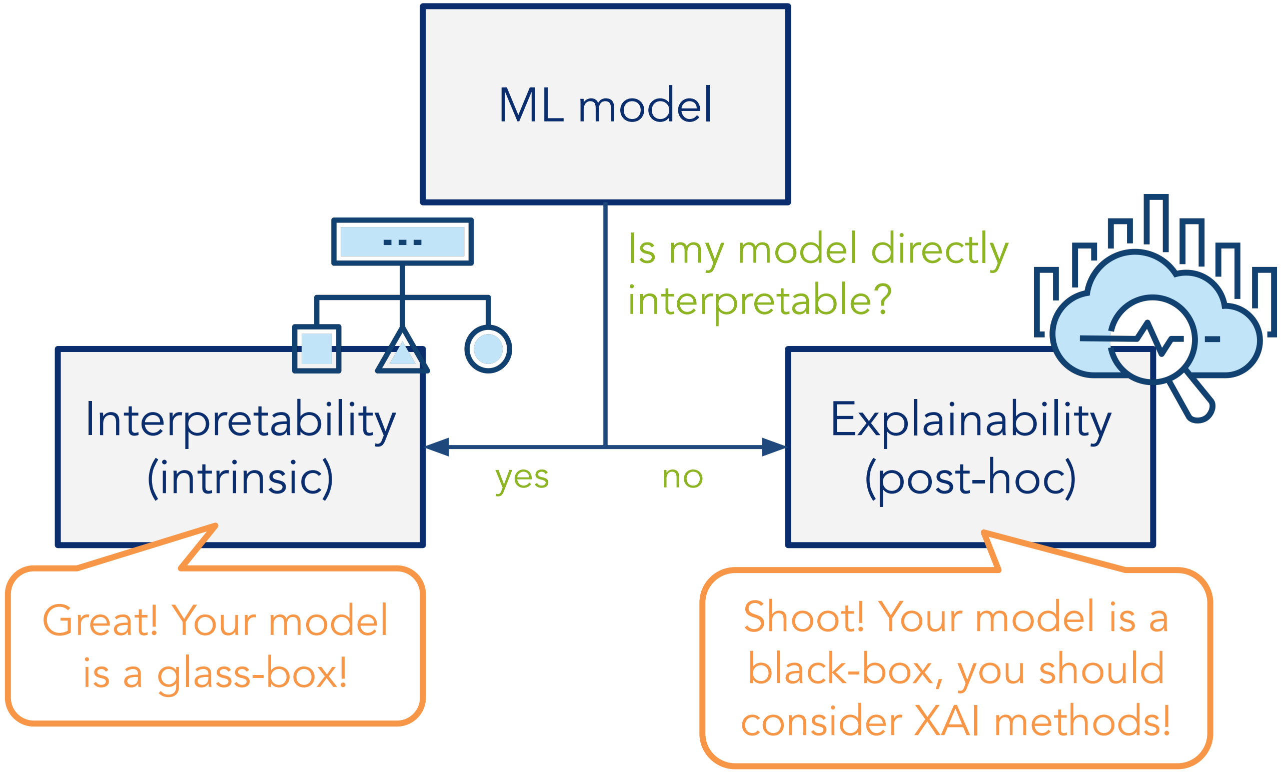
Introduction

Taxonomy of XAI methods

ML model

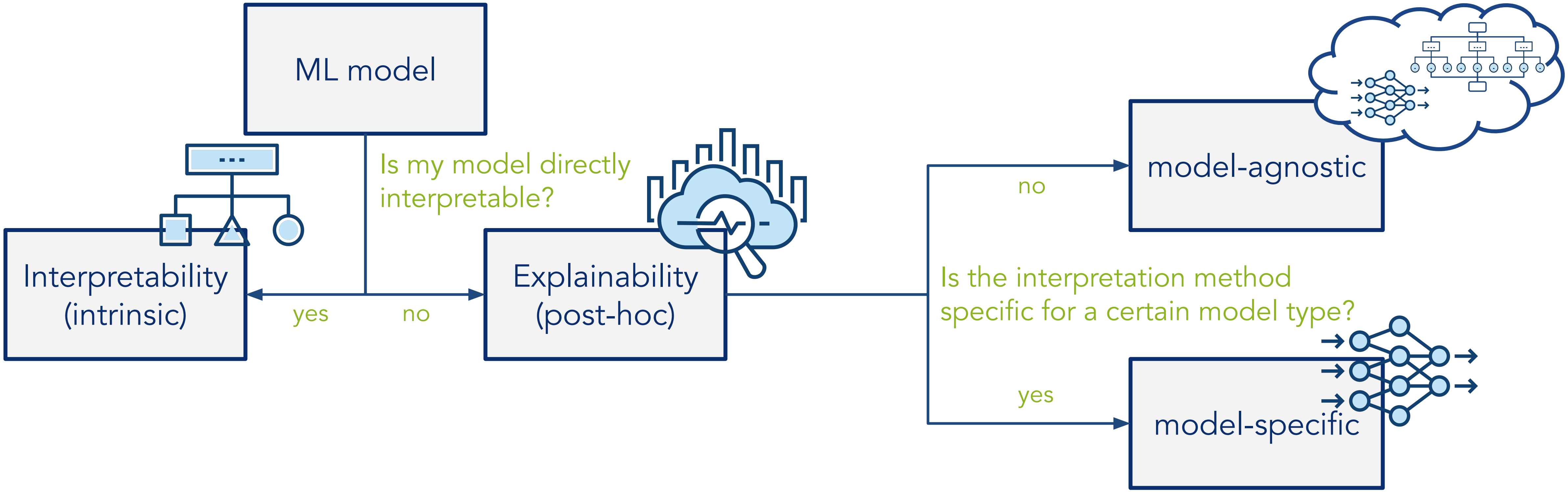
Introduction

Taxonomy of XAI methods



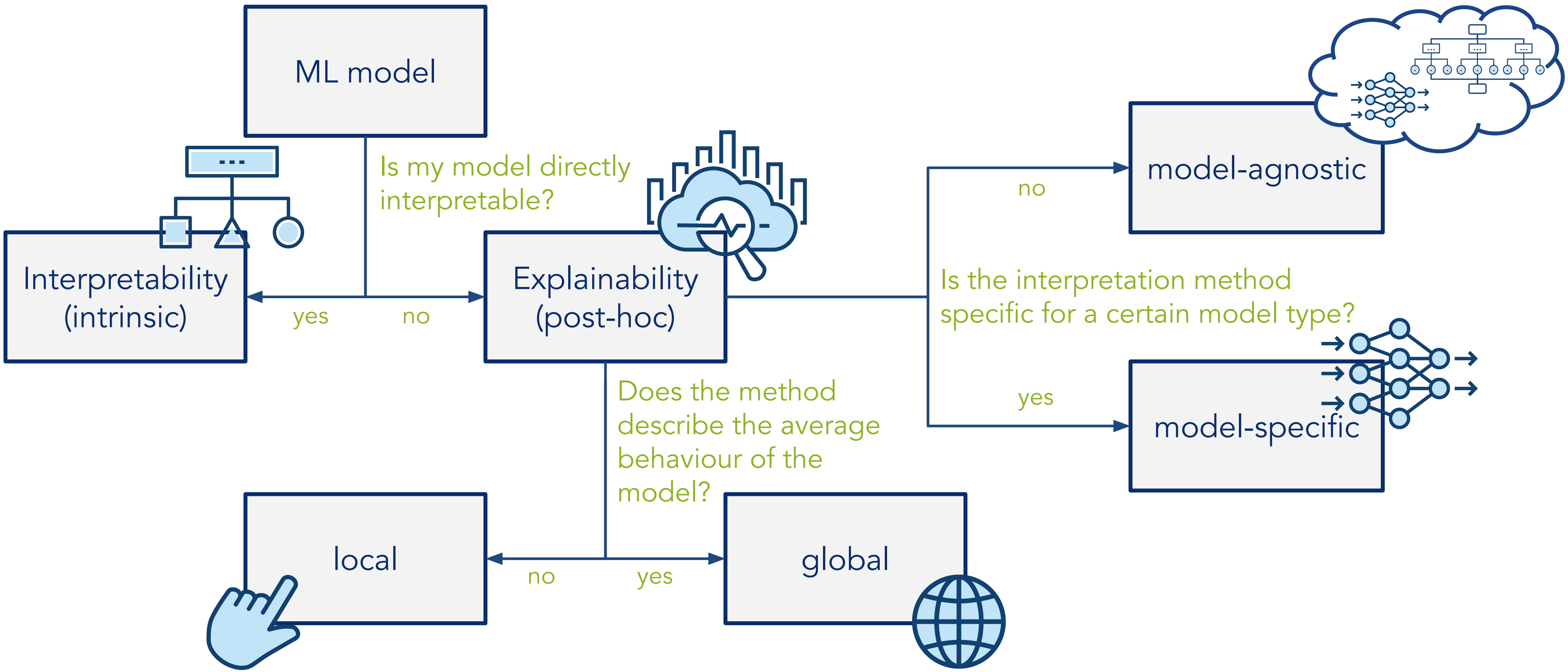
Introduction

Taxonomy of XAI methods



Introduction

Taxonomy of XAI methods



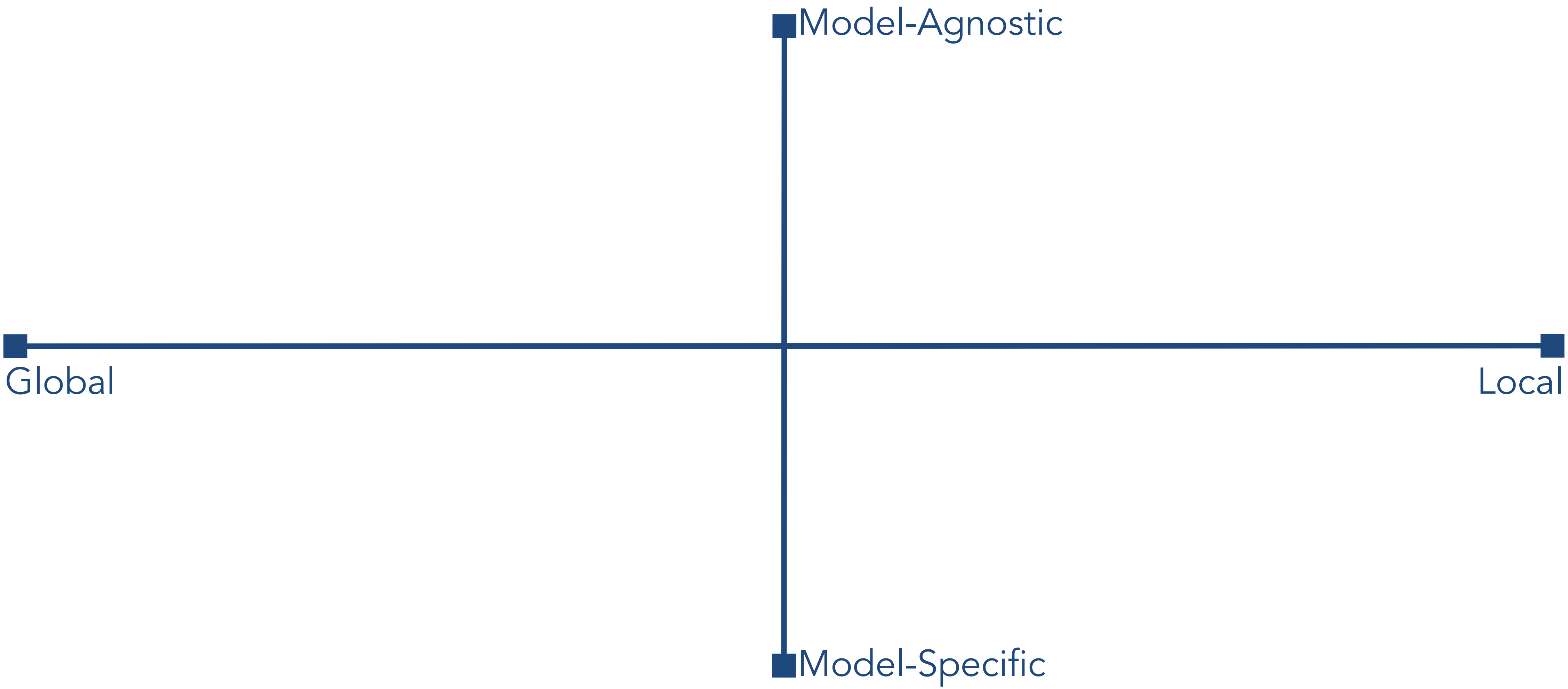
slido

To understand what impact blood pressure has on the survival rate of patient John Doe in a Random Forest model, we need:

ⓘ Start presenting to display the poll results on this slide.

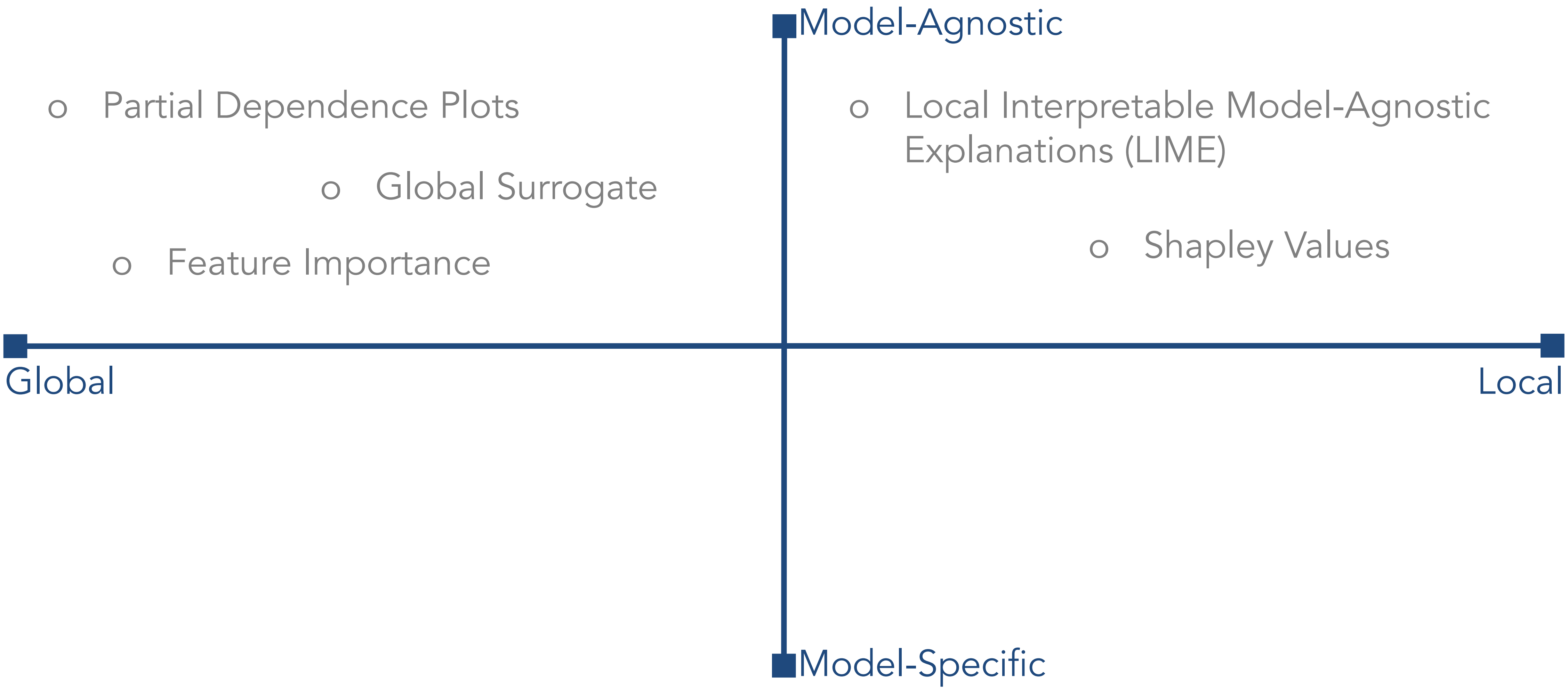
Introduction

Overview on post-hoc methods



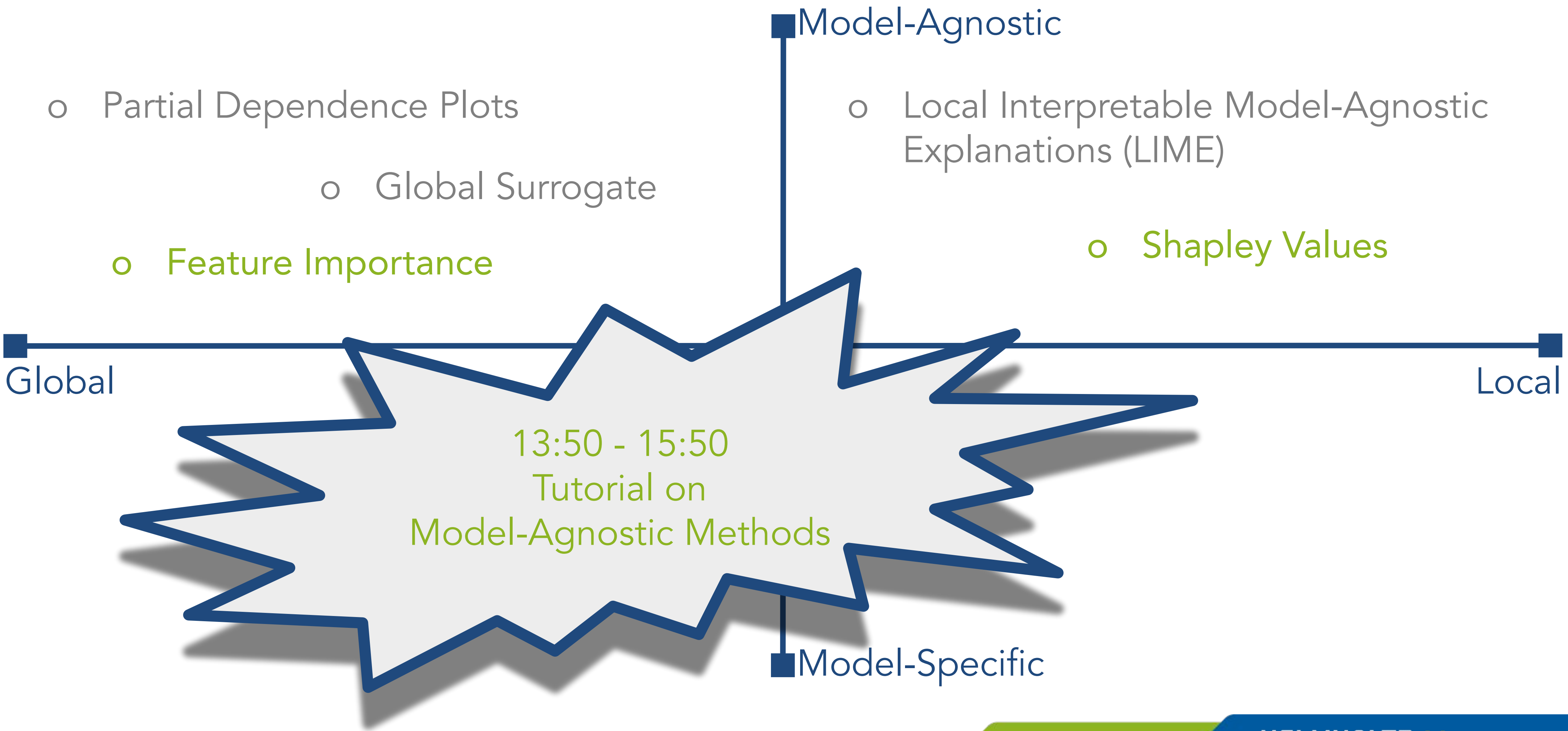
Introduction

Overview on post-hoc methods



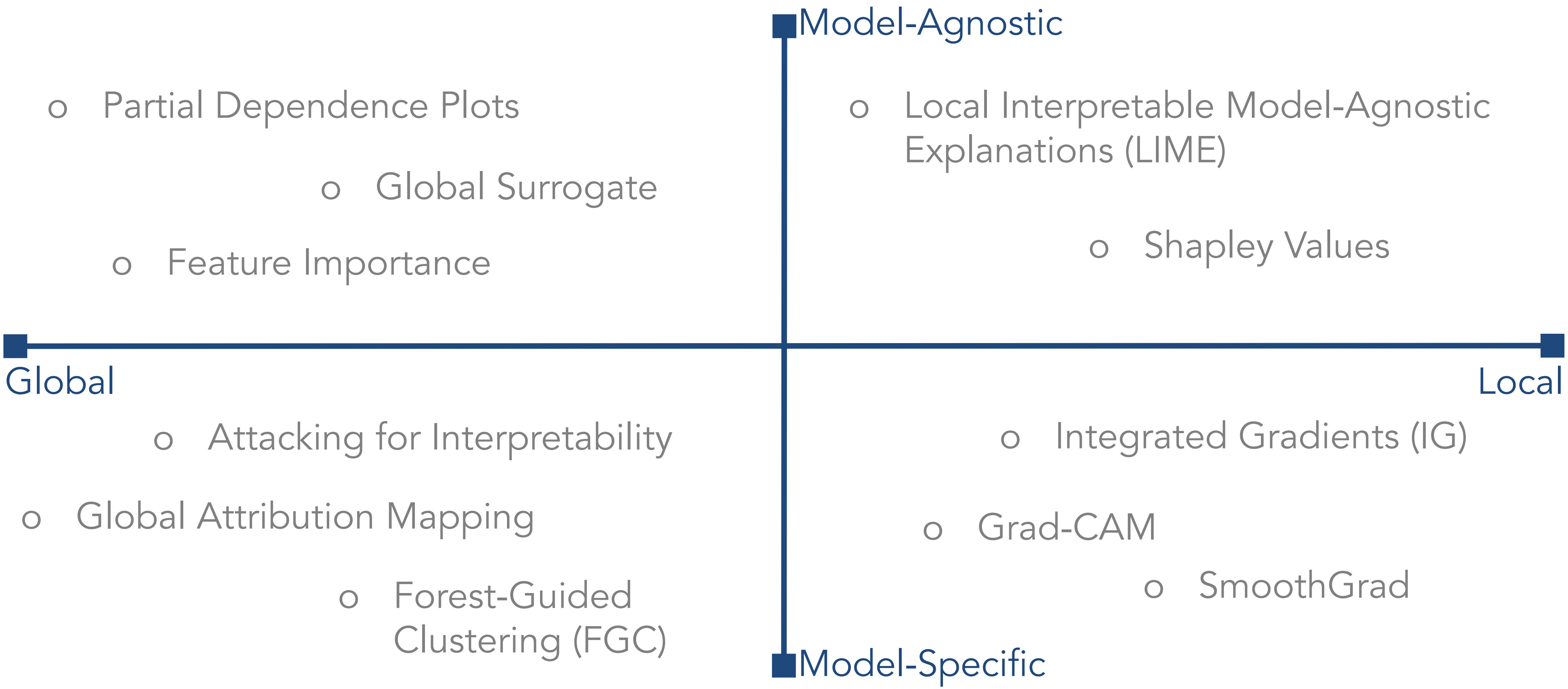
Introduction

Overview on post-hoc methods



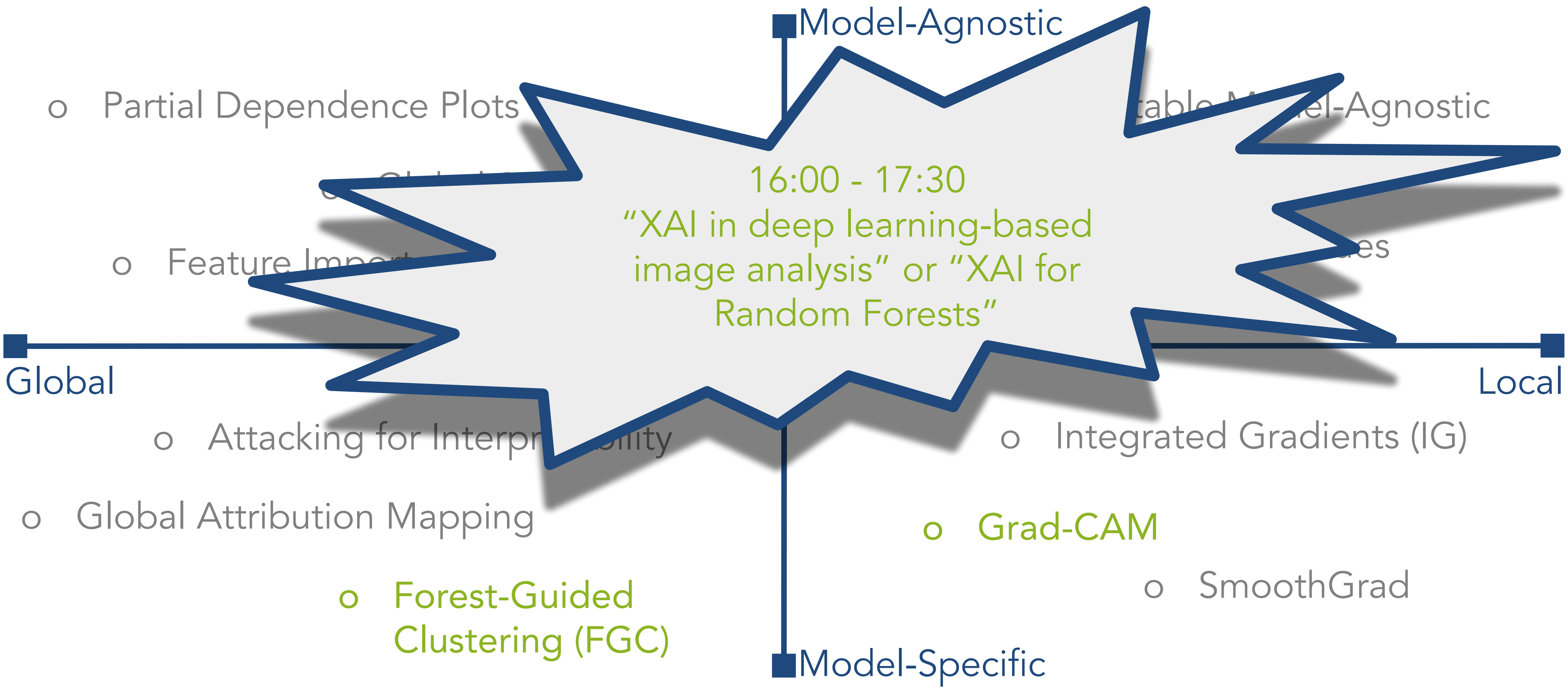
Introduction

Overview on post-hoc methods



Introduction

Overview on post-hoc methods



We will move you now into separate breakout sessions with your tutors!

HAVE FUN WITH THE TUTORIALS!

Who are we?

Helmholtz AI

If you have questions on Helmholtz AI, contact us at:
consultant-helmholtz.ai@helmholtz-muenchen.de

WHAT IS OUR MISSION?

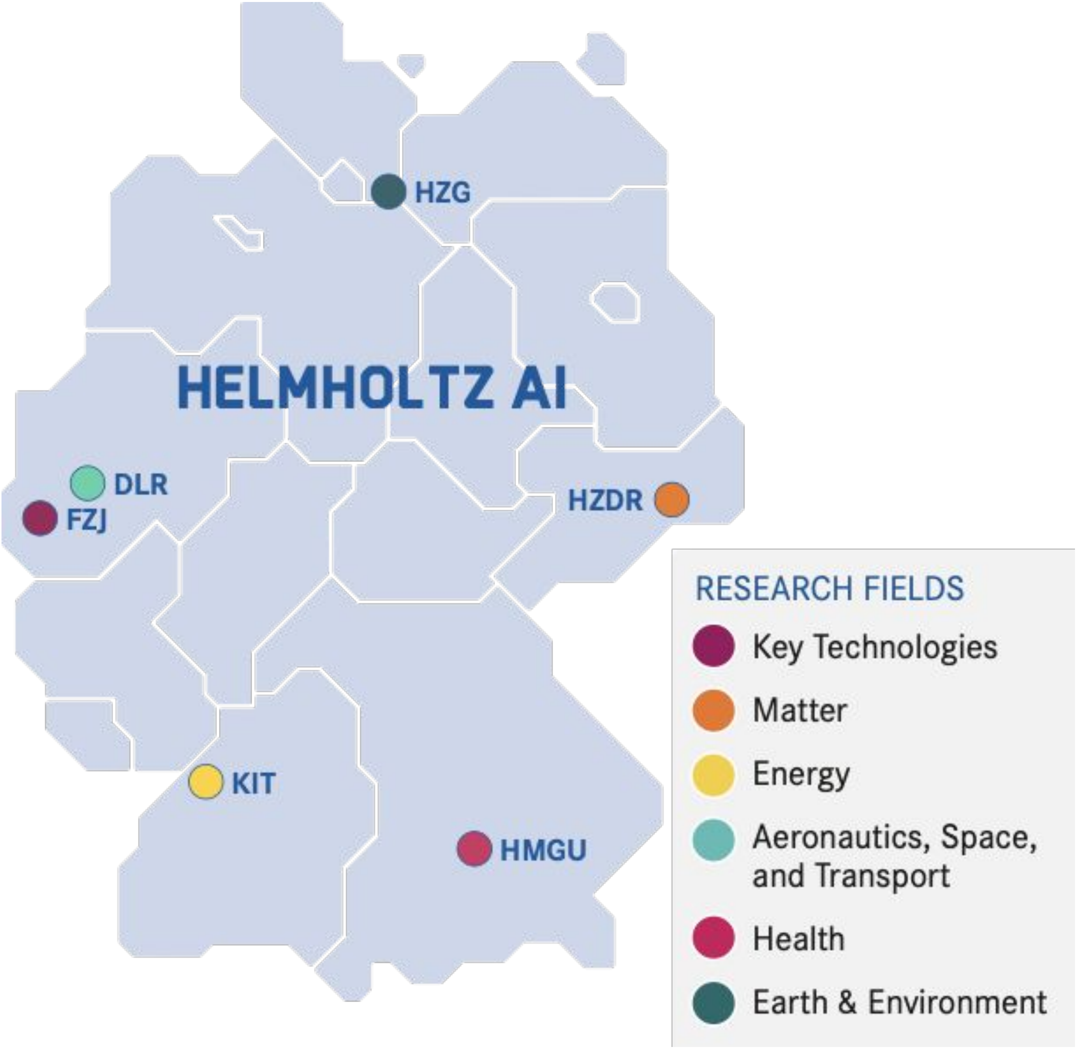
Maximise research impact by democratising access to AI

WHO ARE WE?

Interdisciplinary platform for innovative research in AI

Compiles develops and fosters applied AI methods nationwide across all Helmholtz Centers

Aims to reach international leadership in applied AI

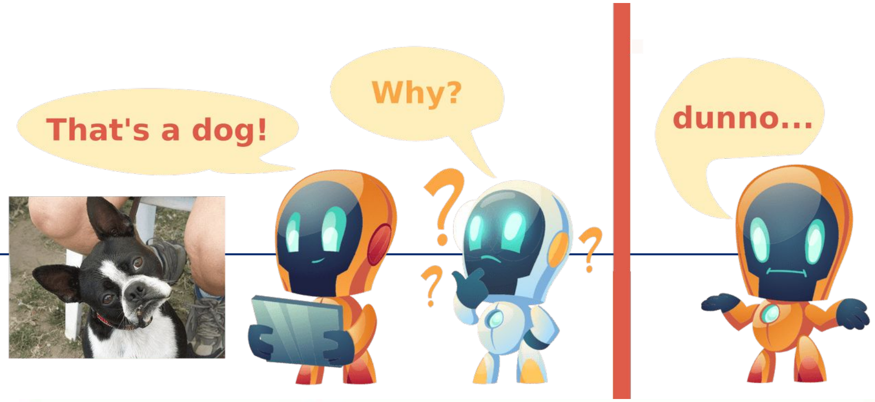


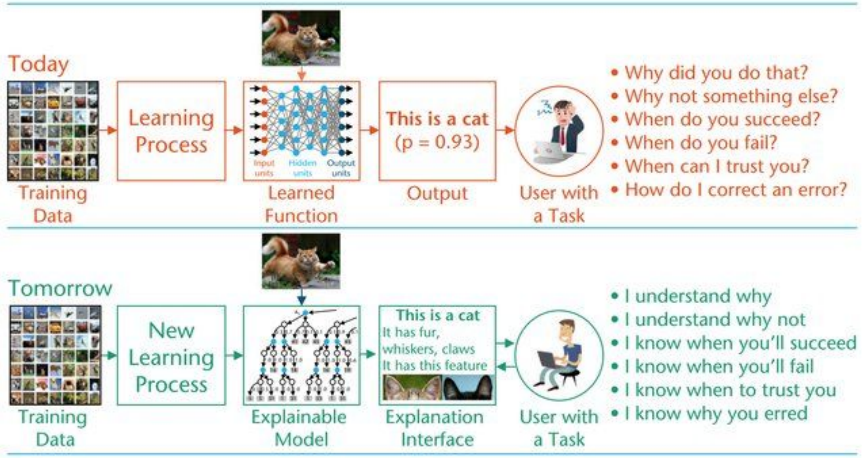
Additional Resources

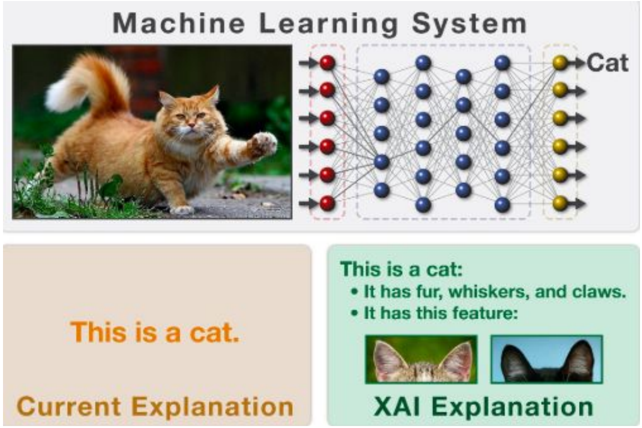
References for figures:
- Im1: https://erdem.pl/2021/10/xai-methods-the-introduction
- Im2:https://www.researchgate.net/publication/351769874_Heading_Toward_Trusted_ATCO-AI_Systems_A_Literature_Review
- Im3: https://twitter.com/Connected_Data/status/918776492292739072/photo/1

Why is it important?

- build trust in AI ti use ML models in sensitive areas (healthcare, legal system), e.g. However, when doctors cannot explain the outcome, they are hesitant to use this technology and act on its recommendations.
  - https://towardsdatascience.com/what-is-explainable-ai-xai-afc56938d513
- motivation why XAI is useful → [predicting pneumonia outcome goes wrong]()
- good into to XAI:
  - https://ambiata.com/blog/2021-04-12-xai-part-1/
  - https://blogs.nvidia.com/blog/2021/05/24/what-is-explainable-ai/
  - https://towardsdatascience.com/explainable-ai-9a9af94931ff
- case studies:
  - https://www.nature.com/articles/s41598-021-02370-4
  - https://arxiv.org/pdf/2010.02006.pdf
- AI in healthcare:
  - http://www.comp.hkbu.edu.hk/~cib/2018/Aug/article1/iib_vol19no1_article1.pdf
  - Round table Interviews: https://www.vanderschaar-lab.com/interpretable-machine-learning/







Introduction

Terminology

How is Interpretability defined?

Introduction

Terminology

„[…] interpretability is the degree to which a human can understand the cause of a decision […]"

— (Miller et al., 2019)

„[…] interpretability is the degree to which a human can consistently predict the model's result […]"

— (Kim et al., 2016)

How is Interpretability defined?

„[…] in machine learning, interpretability is defined as the ability to explain or to provide the meaning in understandable terms to a human […]"

— (Guidotti et al., 2018)

„[…] interpretability is defined as the ability to explain or to provide the meaning in understandable terms to a human […]"

— (Doshi-Velez et al., 2017)